CHAPTER 20

Multivariate analysis: an overview

Chapter contents

20.1	Introduction	610	20.3	Dependency techniques	612
20.2	Selecting a multivariate technique	610	20.4	Interdependency techniques	628

Learning objectives

When you have read this chapter, you should understand:

1. how to classify and select multivariate techniques
2. which kind of questions you can answer by applying the different multivariate techniques
3. the assumptions made when using each of these analysis techniques.

20.1 Introduction

In recent years, multivariate statistical tools have been applied with increasing frequency to research problems. This recognizes that many problems we encounter are more complex than the problems bivariate models can explain. Simultaneously, computer programs have taken advantage of the complex mathematics needed to manage multiple variable relationships. Today, computers with fast processing speeds and versatile software bring these powerful techniques to researchers.

Throughout the functional areas of management, more and more problems are being addressed by considering multiple independent and/or multiple dependent variables. Sales managers base forecasts on various product history variables; marketers consider the complex set of buyer preferences and preferred product options; financial analysts classify levels of credit risk based on a set of predictors; and human resource managers devise future wage and salary compensation plans with multivariate techniques.

Many of the examples presented in this text could be considered multivariate problems. The revenue improvements for a physicians' group that decided to join a different insurance programme were based on multiple factors. In another example, the aviation industry was attempting to control radiation risks for passengers and crew by altering the proximity of air routes to the poles, aircraft shielding, altitude and other variables. The price of investment-grade wine was forecast based on spring and harvest rainfall and growing-season temperatures.

One author defines **multivariate analysis** as 'those statistical techniques that focus upon, and bring out in bold relief, the structure of simultaneous relationships among three or more phenomena'.[1] Our overview of multivariate analysis seeks to illustrate the meaning of this definition while building on your understanding of bivariate statistics from the last few chapters. Several common multivariate techniques and examples are discussed.

Because a complete treatment of this subject would require a thorough consideration of the mathematics, assumptions and diagnostic tools appropriate for each technique, our coverage is necessarily limited. Readers needing greater detail are referred to the 'Recommended further reading' section at the end of the chapter.

20.2 Selecting a multivariate technique

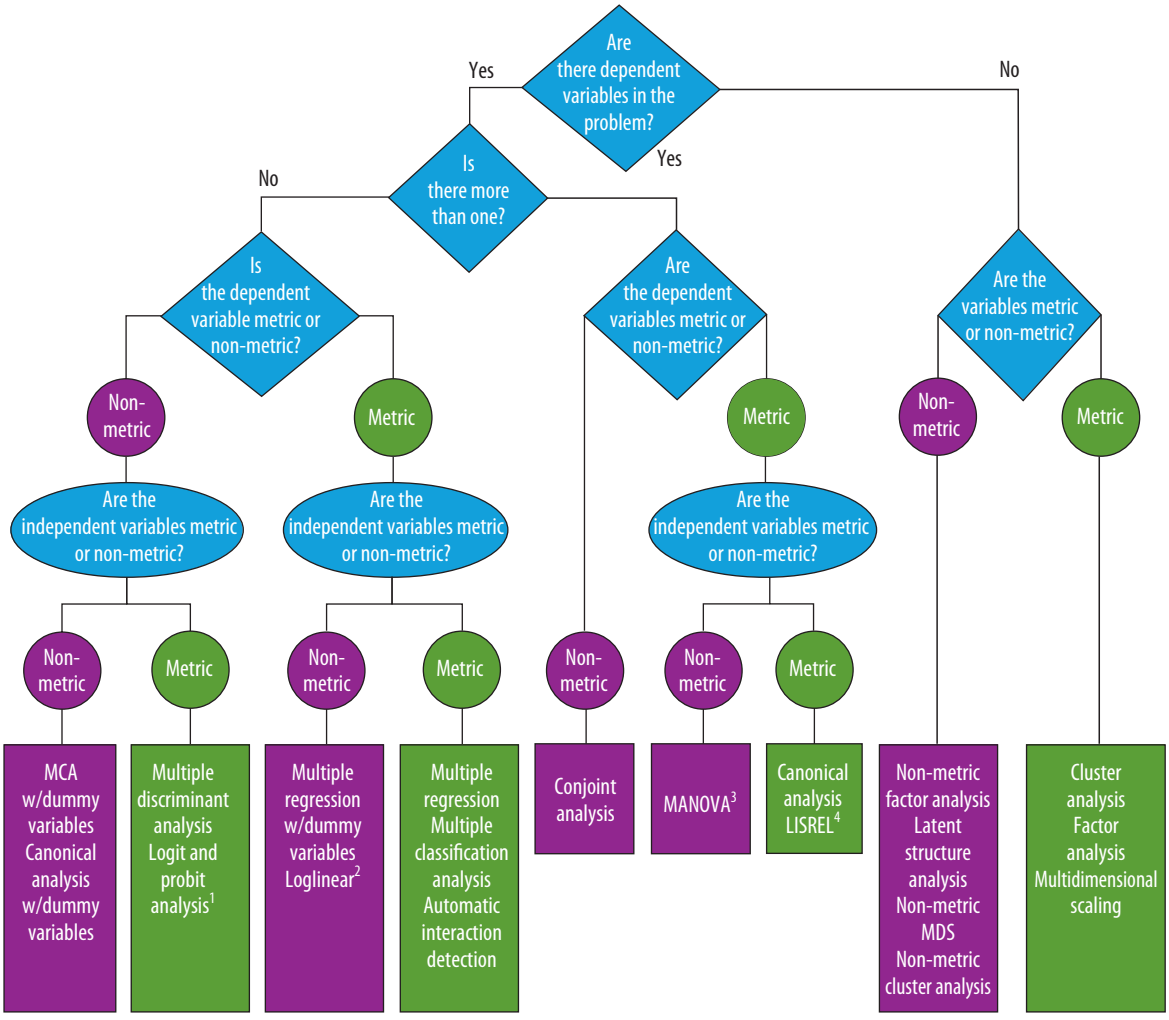
Multivariate techniques may be classified as **dependency techniques** and **interdependency techniques**. Selecting an appropriate technique starts with an understanding of this distinction. If criterion and predictor variables exist in the research question, then we will have an assumption of dependence. Multiple regression, multivariate analysis of variance (MANOVA) and discriminant analysis are techniques where criterion or dependent variables and predictor or independent variables are present. Alternatively, if the variables are interrelated without designating some dependent and others independent, then interdependence of the variables is assumed. Factor analysis, cluster analysis and multidimensional scaling are examples of interdependency techniques.

Exhibit 20.1 provides a diagram as a guide in the selection of techniques. It is also an example to show how you might make a decision.

Every other year since 1978, the Roper organization has tracked public opinion towards business by providing a list of items that are said to be the responsibility of business. The respondents are asked whether business fulfils these responsibilities 'fully', 'fairly well', 'not too well' or 'not at all well'. The following issues make up the list:[2]

- developing new products and services
- producing good-quality products and services
- making products that are safe to use
- hiring minorities
- providing jobs for people
- being good citizens of the communities in which they operate
- paying good salaries and benefits to employees
- charging reasonable prices for goods and services
- keeping profits at reasonable levels

Exhibit 20.1 ***Selecting from the most common multivariate techniques.***



Source: Partially adapted from T.C. Kinnear and J.R. Taylor, 'Multivariate methods in marketing: a further attempt at classification', *Journal of Marketing*, October 1971, p. 57; and J.F. Hair Jr., Rolph E. Anderson, Ronald L. Tatham and Bernie J. Grablowsky, *Multivariate Data Analysis* (Tulsa, OK: Petroleum Publishing Co., 1979), pp. 10–14.

Notes

1 The independent variable is metric only in the sense that a transformed proportion is used.

2 The independent variable is metric only when we consider that the number of cases in the cross-tabulation cell are used to calculate the logs.

3 Factors may be considered non-metric independent variables in that they organize the data into groups. We do not classify MANOVA and other multivariate analysis of variance models.

4 LISREL refers to a linear structural equations model for latent variables. It is a family of models appropriate for confirmatory factor analysis, path analysis, time series analysis, recursive and non-recursive models, and covariance structure models. Because it may handle dependence and interdependence, metric and non-metric, it is arbitrarily placed in this diagram.

- advertising honestly
- paying their fair share
- cleaning up their own air and water pollution.

You have access to data on these items and wish to know if they could be reduced to a smaller set of variables that would account for most of the variation among respondents. In response to the first question in Exhibit 20.1, you correctly determine there are no dependent variables in the dataset. You then check to see if the variables are **metric measures** or **non-metric measures**. In the exhibit, metric refers to ratio and interval measurements, and non-metric refers to data that are nominal and ordinal. Based on the measurement scale, which appears to have

equal intervals, and preliminary findings that show a linear relationship between several variables, you decide that the data are metric. This decision leads you to three options: multidimensional scaling, cluster analysis or factor analysis. Multidimensional scaling develops a geometric picture or map of the locations of some objects relative to others. This map specifies how the objects differ. Cluster analysis identifies homogeneous sub-groups or clusters. Factor analysis looks for patterns among the variables to discover if an underlying combination of the original variables (a factor) can summarize the original set. Based on your research objective, you select factor analysis.

Suppose you are interested in predicting family food expenditures from family income, family size, and whether the family's location is rural or urban. Returning to Exhibit 20.1, you conclude that there is a singular dependent variable, family food expenditures. You decide this variable is metric since dollars are measured on a ratio scale. The independent variables, income and family size, also meet the criteria for metric data. However, you are not sure about the location variable since it appears to be a dichotomous nominal variable. According to the figure, your choices are automatic interaction detection (AID), multiple classification analysis (MCA) and multiple regression. AID was designed to locate the most important interaction effects and typically uses numerous independent variables in its sequential partitioning procedure. MCA handles weak predictors (including nominal variables), correlated predictors and non-linear relationships. Multiple regression is the extension of bivariate regression. You believe that your data exceed the assumptions for the first two techniques and that by treating the nominal variable's values as 0 or 1, you could use it as an independent variable in a multiple regression model. You prefer this to losing information from the other two variables – a certainty if you reduce them to non-metric data.

In the next two sections, we extend this discussion as we illustrate dependency and interdependency techniques.

20.3 Dependency techniques

Multiple regression

Multiple regression is used as a descriptive tool in three types of situation. First, it is often used to develop a self-weighting estimating equation by which to predict values for a criterion variable (DV) from the values for several predictor variables (IVs). Thus, we might try to predict company performance (measured as ROI – return on investment) on the basis of market share, number of approved patents in the last three years, degree of internationalization and a time factor. Another prediction study might be one in which we estimate the chances of becoming an entrepreneur from the variables work experience, having self-employed parents, availability of financial funds and the number of people one knows.

Second, a descriptive application of multiple regression calls for controlling for confounding variables to better evaluate the contribution of other variables. For example, one might wish to control the brand of a product and the store in which it is bought to study the effects of price as an indicator of product quality.[3] A third use of multiple regression is to test and explain causal theories. In this approach, often referred to as **path analysis**, regression is used to describe an entire structure of linkages that have been advanced from a causal theory.[4] Finally, in academic research, multiple regression is often used as an inference tool to test hypotheses and to estimate population values.

Method

Multiple regression is an extension of the bivariate linear regression presented in Chapter 19. The terms defined in that chapter will not be repeated here. Although **dummy variables** (nominal variables coded 0, 1) may be used, all other variables must be interval or ratio. Nominal variables with more than two values may also be used if they are transformed to a set of dummy variables. For example, to include the multi-nominal variable UK regions with the four values 'England', 'Northern Ireland', 'Scotland' and 'Wales', you would construct $n - 1$, that is $4 - 1 =$ three, dummy variables. The dummy variable 'England' would be coded 1 if the participant lives in England and 0 otherwise; the dummy variable 'Northern Ireland' would be coded 1 if the participant lives in Northern Ireland and 0 otherwise; the dummy variable 'Scotland' would be coded 1 if the participant lives in Scotland and 0 otherwise. You would not include a dummy variable 'Wales', as a participant living in Wales can be identified by looking at

the three dummy variables 'England', 'Northern Ireland' and 'Scotland' – these three variables would be coded 0 for a Welsh participant. The generalized equation is:

$$Y = \beta + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_n X_n + \varepsilon$$

where

β_1 = A constant, the value of Y when all X values are zero.
β_2 = The slope of the regression surface or the response surface. The β represents the regression coefficient associated with each X_i.
ε = An error term, normally distributed about a mean of 0. For purposes of computation, the ε is assumed to be 0.

The regression coefficients are stated either in raw score units (the actual X values) or as **standardized coefficients** (X values restated in terms of their standard deviations). In either case, the value of the regression coefficient states the amount that Y varies with each unit change of the associated X variable when the effects of all other X variables are being held constant. When the regression coefficients are standardized, they are called **beta weights** (β), and their values indicate the relative importance of the associated X values, particularly when the predictors are unrelated. For example, in an equation where $\beta_1 = .60$ and $\beta_2 = .20$, one concludes that X_1 has three times the influence on Y as does X_2.

Example

In a Research Methods in Real Life box later in this chapter, we describe an e-business that uses multivariate approaches to understand its target market in the global 'hybrid mail' business. SuperLetter's basic service enables users to create a document on any PC and send it in a secure encrypted mode over the Internet to a distant international terminal near the addressee, where it will be printed, processed and delivered via the local postal service. Spread like a 'fishing net' over the world's major commercial markets, the network connects corresponding parties, linking the world's 'wired' with its 'non-wired'.

We use multiple regression in this example to evaluate the key drivers of customer usage for hybrid mail. Among the explanatory variables are customer perceptions of: (i) cost/speed valuation, (ii) security (limits on changing, editing or forwarding a document, and document privacy), (iii) reliability, (iv) receiver technology (hard copy for receivers with no email or fax access), and (v) impact/emotional value (reducing email spam clutter and official/important appearance). We have chosen the first three variables, all measured on five-point scales, for this equation:

Y = customer usage
X_1 = cost/speed valuation
X_2 = security
X_3 = reliability

SPSS computed the model and the regression coefficients. Most statistical packages provide various methods for selecting variables for the equation. The equation can be built with all variables or specific combinations or you can select a method that sequentially adds or removes variables (forward selection, backward elimination and stepwise selection). **Forward selection** starts with the constant and adds variables that result in the largest R^2 increase. **Backward elimination** begins with a model containing all independent variables and removes the variable that changes R^2 the least. **Stepwise selection**, the most popular method, combines forward and backward sequential approaches. The independent variable that contributes the most to explaining the dependent variable is added first. Subsequent variables are included based on their incremental contribution over the first variable and whether they meet the criterion for entering the equation (e.g. a significance level of .01). Variables may be removed at each step if they meet the removal criterion, which is a larger significance level than for entry.

The standard elements of a stepwise output are shown in Exhibit 20.2. In the upper portion of the exhibit there are three models. In Model 1, cost/speed is the first variable to enter the equation. This model consists of the constant and the variable cost/speed. Model 2 adds the security variable to cost/speed. Model 3 consists of all three independent variables. In the summary statistics for Model 1, you see that cost/speed explains 77 per cent of customer usage (see the R^2 column). This is increased by 8 per cent in Model 2 when security is added (see R^2 Change column). When reliability is added in Model 3, accounting for only 2 per cent, 87 per cent of customer usage is explained.

Exhibit 20.2 ***Multiple regression analysis of hybrid mail customer usage, cost/speed evaluation, security and reliability.***

Model summary

Model	R	Adjusted R^2	R^2	Std. error of the estimate	R^2 change	Change statistics			
						F change	*d.f.*1	*d.f.*2	Sig. *F* change
1	.879	.772	.771	.658	.772	612.696	1	181	.000
2	.925	.855	.854	.526	.083	103.677	2	180	.000
3	.935	.873	.871	.493	.018	25.597	3	179	.000

1 Predictors: (constant), cost/speed.
2 Predictors: (constant), cost/speed, security.
3 Predictors: (constant), cost/speed, security, reliability.

Coefficient

Model		Unstandardized coefficients	Std. error	Standardized coefficients	*t*	Sig.	Collinearity statistics
		B		Beta			VIF
1	(Constant)	.579	.151		3.834	.000	
	Cost/Speed	.857	.035	.879	24.753	.000	1.000
2	(Constant)	9.501E-02	.130	.733	.464		
	Cost/Speed	.537	.042	.551	12.842	.000	2.289
	Security	.428	.042	.437	10.182	.000	2.289
3	(Constant)	–9.326E-02	.127	–.734	.464		
	Cost/Speed	.448	.043	.460	10.428	.000	2.748
	Security	.315	.045	.321	6.948	.000	3.025
	Reliability	.254	.050	.236	5.059	.000	3.067

Note: Dependent variable: customer usage.

A critical side note has to be placed with regard to the use of this method selecting variables to be included. The selection is based on a statistical criterion (change in R^2), that is, how much the added independent variable adds to the explained variance of the depended variable. However, a large change in R^2 does not necessarily inform us about the relevance of the variable or the goodness of our theoretical model. First, an independent variable that is rather similar to the dependent variable usually explains an important portion of the dependent variable's variance, but might hardly explain it, that is the phenomenon under investigation. For example, if you want to explain why the sales of breweries differ (dependent variable), independent variables like the number of customer or sales in the previous year will typically be highly significant and explain a large part of the variations in sales. However, they hardly add to our understanding of why some breweries have large sales and others small sales. Second, we obtain sometimes significant coefficients with a sign that contradicts our expectations; for example, although we expected on theoretical grounds that more heterogeneous teams are more creative, we observe that the coefficient of the independent variable 'team heterogeneity' is significantly negative, thus more homogeneous teams are more creative. Including the variable 'team heterogeneity' will increase our R^2. An interpretation of the R^2 in the sense that our theoretical considerations are supported by the analysis is, however, ill-founded, because the R^2 is partly based on an effect contradicting our theoretical considerations.

The other reported statistics have the following interpretations:

1 Adjusted *R* squared for Model 3 = .871. R^2 is adjusted to reflect the model's goodness of fit for the population. The net effect of this adjustment is to reduce the R^2 from .873 to .871, thereby making it comparable to other R^2's from equations with a different number of independent variables.

2 Standard error of Model 3 = .4937. This is the standard deviation of actual values of *Y* about the regression line of estimated *Y* values.
3 Analysis of variance measures, whether or not the equation represents a set of regression coefficients that, in total, are statistically significant from zero. The critical value for *F* is found in Appendix E, Exhibit E.9, with degrees of freedom for the numerator equalling *k*, the number of independent variables, and for the denominator, $n - k - 1$, where *n* for Model 3 is 183 observations. Thus, *d.f.* = (3, 179). The equation is statistically significant at less than the .05 level of significance (see the column labelled 'Significant *F* Change').
4 Regression coefficients for all three models are shown in the lower table of Exhibit 20.2. The column headed 'B' shows the unstandardized regression coefficients for the equation. The equation may now be constructed as:

$$Y = -.093 + .448X_1 + .315X_2 + .254X_3$$

5 The column headed 'Beta' gives the regression coefficients expressed in standardized form. When these are used, the regression *Y* intercept is zero. Standardized coefficients are useful when the variables are measured on different scales. The beta coefficients also show the relative contribution of the three independent variables to the explanatory power of this equation. The cost/speed valuation variable explains more than either of the other two variables.
6 Standard error is a measure of the sampling variability of each regression coefficient.
7 The column headed '*t*' measures the statistical significance of each of the regression coefficients.

Again compare these to the table of *t* values in Appendix E, Exhibit E.2, using degrees of freedom for one independent variable. All three regression coefficients are judged to be significantly different from zero. Therefore, the regression equation shows the relationship between the dependent variable, customer usage of hybrid mail, and three independent variables: cost/speed, security and reliability. The regression coefficients are both individually and jointly statistically significant. The independent variable cost/speed influences customer usage the most, followed by security and then reliability.

Collinearity or **multicollinearity** – the situation where two or more of the independent variables are highly correlated – can have damaging effects on multiple regression. When this condition exists, the estimated regression coefficients can fluctuate widely from sample to sample, making it risky to interpret the coefficients as an indicator of the relative importance of predictor variables. Just how high can acceptable correlations be between independent variables? There is no definitive answer, but correlations at a .80 or greater level should be dealt with in one of two ways:

1 If the two variables are theoretically distinct, choose one of the variables and delete the other.
2 If the two variables are theoretically related, create a new variable that is a composite of the highly correlated variables and use this new variable in place of its components.

Making this decision with a correlation matrix alone is not always advisable. In the example just presented, Exhibit 20.2 contains a column labelled 'Collinearity statistics' that shows a variable inflation factor (VIF) index. This is a measure of the effect of the other independent variables on a regression coefficient. Large values, usually 10.0 or more, suggest collinearity or multicollinearity. With the three predictors in the hybrid mail example, multicollinearity is not a problem.

Another difficulty with regression occurs when researchers fail to evaluate the equation with data beyond those used originally to calculate it. A practical solution is to set aside a portion of the data (from a fourth to a third) and use only the remainder to compute the estimating equation. This is called a **holdout sample**. One then uses the equation on the holdout data to calculate an R^2. This can then be compared to the original R^2 to determine how well the equation predicts beyond its database.

SPSS reference

How to conduct a multiple regression analysis in SPSS is documented in Chapter 13 of Pallant (2013).

Discriminant analysis

Researchers often wish to classify people or objects into two or more groups. One might need to classify persons as either buyers or non-buyers, good or bad credit risks, or superior, average or poor performers in some activity. The objective is to establish a procedure to find the predictors that best classify subjects.

Method

Discriminant analysis joins a nominally scaled criterion or dependent variable with one or more independent variables that are interval- or ratio-scaled. Once the discriminant equation is found, it can be used to predict the classification of a new observation. This is done by calculating a linear function of the form:

$$D_i = d_0 + d_1X_1 + d_2X_2 + \ldots + d_pX_p$$

where

D_i is the score on discriminant function i.
The d_is are weighting coefficients; d_0 is a constant.
The Xs are the values of the discriminating variables used in the analysis.

A single discriminant equation is required if the categorization calls for two groups. If three groups are involved in the classification, it requires two discriminant equations. If more categories are called for in the dependent variable, it is necessary to calculate a separate discriminant function for each pair of classifications in the criterion group.

While the most common use for discriminant analysis is to classify persons or objects into various groups, it can also be used to analyse known groups to determine the relative influence of specific factors for deciding into which group various cases fall. Assume we have supervisory ratings that enable us to classify administrators as successful or unsuccessful on administrative performance. We might also be able to secure test results on three measures: ability to work with others (X_1), motivation for administrative work (X_2) and general professional skill (X_3). Suppose the discriminant equation is:

$$D = .06X_1 + .45X_2 + .30X_3$$

Since discriminant analysis uses standardized values for the discriminant variables, we conclude from the coefficients that ability to work with others is less important than the other two in classifying administrators.[5]

Exhibit 20.3 ***Discriminant analysis classification results on MBA hires at Dean Merrill.***

		Predicted group membership	
Actual group		0	1
Unsuccessful 0	15	13	2
		86.7%	13.3%
Successful 1	15	3	12
		20.0%	80.05%

	Unstandardized	Standardized
X_1	.36084	.65927
X_2	2.61202	.57958
X_3	.53028	.97505
(constant)	12.89685	
	5	

Example

An illustration of the method takes us back to the problem presented in the last chapter where Dean Merrill, a brokerage firm, is hiring MBAs for its account executives programme. Over the years the firm has had indifferent success with the selection process. You are asked to develop a procedure to improve it. It appears that discriminant analysis is a perfect technique. You begin by gathering data on 30 MBAs who have been hired in recent years; 15 of these have been successful employees while the other 15 have been unsatisfactory. The personnel files provide the following information that can be used to conduct the analysis:

X_1 = Years of prior work experience
X_2 = GPA in graduate programme
X_3 = Employment test scores

An algorithm determines how well these three independent variables will correctly classify those who are judged successful from those judged unsuccessful. The classification results are shown in Exhibit 20.3. This indicates that 25 of the 30 ($30 - 3 - 2 = 25$) cases have been correctly classified using these three variables.

The standardized and unstandardized discriminant function coefficients are shown in the second panel of Exhibit 20.3. These results indicate that X_3 (the employment test) has the greatest discriminating power. Several significance tests are also computed.

One, Wilk's lambda, has a chi-square transformation for testing the significance of the discriminant function. It indicates that the equation is statistically significant at the $\alpha = .0004$ level. Using the discriminant equation:

$$D = .659X_1 + .580X_2 + .975X_3$$

you can now predict whether future candidates are likely to be successful account executives.

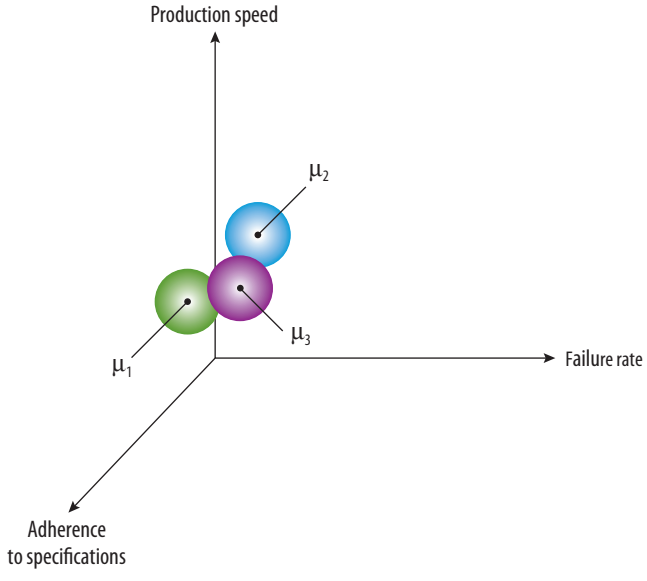
MANOVA

Multivariate analysis of variance (MANOVA) is a commonly used multivariate technique. MANOVA assesses the relationship between two or more dependent variables and classificatory variables or factors. In business research, MANOVA can be used to test differences among samples of employees, customers, manufactured items, production parts, and so on.

Method

MANOVA is similar to the univariate ANOVA described earlier, with the added ability to handle several dependent variables. If ANOVA is applied consecutively to a set of interrelated dependent variables, erroneous conclusions may result. MANOVA can correct this by simultaneously testing all the variables and their interrelationships. MANOVA employs sums-of-squares and cross-products (SSCP) matrices to test for differences among groups. The variance between groups is determined by partitioning the total SSCP matrix and testing for significance. The F ratio, generalized to a ratio of the within-group variance and total-group variance matrices, tests for equality among treatment groups. MANOVA examines similarities and differences among the multivariate mean scores of several populations. The null hypothesis for MANOVA is that all of the **centroids** (multivariate means) are equal, $H_0: \mu_1 = \mu_2 = \mu_3 = \ldots \mu_n$. The alternative hypothesis is that the vectors of centroids are unequal, $H_A: \mu_1 \neq \mu_2 \neq \mu_3 \neq \ldots \mu_n$. Exhibit 20.4 shows graphically three populations whose centroids are unequal, allowing the researcher to reject the null hypothesis. When the null hypothesis is rejected, additional tests are done to better understand the data. Several alternatives may be considered:

Exhibit 20.4 MANOVA techniques show these three centroids to be unequal in the DKD study.



1. univariate F tests can be run on the dependent variables
2. simultaneous confidence intervals can be produced for each variable
3. stepdown analysis, like stepwise regression, can be run by computing F values successively; each value is computed after the effects of the previous dependent variable are eliminated
4. multiple discriminant analysis can be used on the SSCP matrices; this aids in the discovery of which variables contribute to the MANOVA's significance.[6]

Example

To illustrate, take a look at DKD, a firm that manufactures LCD displays. The plant manager is concerned about the quality of the displays coming off the manufacturing line. Two measures are used to assess quality in this

example: adherence to product specifications and time before failure. Measured on a 0–100 scale with 100 meeting all product specifications, the specification variable is averaging approximately 90. The mean time before failure is calculated in weeks; it is approximately 27,000 hours, or 159 weeks or 3 years of continuous operation.

The plant manager asks the industrial engineering department to devise a modified manufacturing procedure that will improve the quality measures but not change the production rate significantly. A new method is designed that includes more efficient parts handling and 'burn-in' time where displays are powered up and run at high temperatures.

Engineering takes a sample of 15 displays made with the old manufacturing method and 15 made with the new method. The displays are measured for their adherence to product specifications and are stress-tested to determine their time before failure. The stress test uses accelerated running conditions and adverse environmental conditions to simulate years of use in a short time.

***Exhibit 20.5** MANOVA cell means and standard deviation in DKD study.*

VARIABLE	FACTOR	LEVEL	MEAN	STD. DEV.
Failure	Method	1	158.867	4.998
	Method	2	181.067	5.994
	For entire sample		169.967	12.524
Specifications	Method	1	89.800	2.077
	Method	2	94.800	2.178
	For entire sample		92.300	3.292
Speed	Method	1	2.126	.061
	Method	2	2.599	.068
	For entire sample		2.362	.249

***Exhibit 20.6** MANOVA homogeneity of variance tests in the DKD study.*

VARIABLE	TEST	RESULTS
Failure		
	Cochran's *C* (14,2) =	.58954, *P* = .506 (approx.)
	Bartlett-Box *F* (1,2352) =	.44347, *P* = .506
Specifications		
	Cochran's *C* (14,2) =	.52366, *P* = .862 (approx.)
	Bartlett-Box *F* (1,2352) =	.03029, *P* = .862
Speed		
	Cochran's *C* (14,2) =	.55526, *P* = .684 (approx.)
	Bartlett-Box *F* (1,2352) =	.16608, *P* = .684
Multivariate Test for Homogeneity of Dispersion Matrices		
	Box's *M* =	6.07877
	F with (6,5680) *DF* =	.89446, *P* = .498 (approx.)
	Chi-Square with 6 DF =	5.37320, *P* = .497 (approx.)

Exhibit 20.5 shows the mean and standard deviation of the dependent variables (failure, specifications and manufacturing speed) for each level of method.[7] Method 1 represents the current manufacturing process and Method 2 is the new process. The new method extended the time before failure to 181 weeks, compared to 159 weeks for the existing method. The adherence to specifications is also improved, up to 95 from 90. But the manufacturing speed is slower by approximately 30 minutes (.473 hour).

We have used diagnostics to check the assumptions of MANOVA except for equality of variance. Both levels of the manufacturing method variable produce a matrix and the equality of these two matrices must be determined. Exhibit 20.6 contains homogeneity of variance tests for separate dependent variables and a multivariate test. The former are known as univariate tests. The multivariate test is a comparable version that tests the variables simultaneously to determine whether MANOVA should proceed.

The significance levels of Cochran's *C* and Bartlett-Box *F* do not allow us to reject any of the tests for the dependent variables considered separately. This means that the two methods have equal variances in each dependent variable. This fulfils the univariate assumptions for homogeneity of variance. We then consider the variances and covariances simultaneously with Box's M, also found in Exhibit 20.6. Again, we are unable to reject the homogeneity of variance assumption regarding the matrices. This satisfies the multivariate assumptions.

When MANOVA is applied properly, the dependent variables are correlated. If the dependent variables are unrelated, there would be no necessity for a multivariate test, and we could use separate *F*-tests for failure, specifications and speed much like the ANOVAs in Chapter 18. Bartlett's test of

sphericity helps us to decide if we should continue analysing MANOVA results or return to separate univariate tests. In Exhibit 20.7, we look for a determinant value that is close to 0. This implies that one or more dependent variables is a linear function of another. The determinant has a chi-square transformation that simplifies testing for statistical significance. Since the observed significance is below that set for the model ($\alpha = .05$), we are able to reject the null hypothesis and conclude that there are dependencies among the failure, specifications and speed variables.

Exhibit 20.7 ***Bartlett's test of sphericity in the DKD study.***

Statistics for WITHIN CELLS correlations	
Log (Determinant) =	−3.92663
Bartlett's test of sphericity =	102.74687 with 3 D.F.
Significance =	.000
F(max) criterion =	7354.80161 with (3,28) D.F.

We now move to the test of equality of means that considers the three dependent variables for the two levels of manufacturing method. This test is analogous to a *t*-test or an *F*-test for multivariate data. The SSCP matrices are used. Exhibit 20.8 shows three tests, including the Hotelling T^2. All the tests provided are compared to the *F* distribution for interpretation. Since the observed significance level is less than $\alpha = .05$ for the T^2-test, we reject the null hypothesis that said methods 1 and 2 provide equal results with respect to failure, specifications and speed. Similar results are obtained from the Pillai trace and Wilk's statistic.

Exhibit 20.8 ***Multivariate tests of significance in the DKD study.***

Multivariate Tests of Significance (*S* = 1, *M* = 1/2, *N* = 12)					
Test name	Value	Exact *F* hypoth.	*d.f.*	Error *d.f.*	Sig. of *F*
Hotelling	5133492	444.90268	3.00	26.00	.000
Pillai	.98089	444.90268	3.00	26.00	.000
Wilk	.02011	444.90268	3.00	26.00	.000

Finally, to detect where the differences lie, we can examine the results of univariate *F* tests in Exhibit 20.9. Since there are only two methods, the *F* is equivalent to t^2 for a two-sample *t*-test. The significance levels for these tests do not reflect that several comparisons are being made, and we should use them principally for diagnostic purposes. This is similar to problems that require the use of multiple comparison tests in univariate analysis of variance. Note, however, that there are statistically significant differences in all three dependent variables resulting from the new manufacturing method. Techniques for further analysis of MANOVA results were listed at the beginning of this section.

Exhibit 20.9 ***Univariate tests of significance in the DKD study.***

Univariate *F*-tests with (1,28) *d.f.*						
Variable	Hypoth. *SS*	Error *SS*	Hypoth. *MS*	Error *MS*	*F*	Sig. of *F*
Failure	3696.30000	852.66667	3696.30000	30.45238	121.37967	.000
Specs	187.50000	126.80000	187.50000	4.52857	41.40379	.000
Speed	1.67560	.11593	1.67560	.00414	404.68856	.000

SPSS reference

If you would like to replicate the multi-variate Anova shown here in SPSS yourself, see Chapter 21 of Pallant (2013).

LISREL[8]

First developed by Karl Jöreskog in 1973, **LISREL** (an acronym for LInear Structural RELationships) is still a commonly accepted term for referring to both the software program and the general statistical method for modelling

the analysis of covariance structures. LISREL is a powerful alternative to other multivariate techniques, such as multiple regression, MANOVA and canonical analysis – which are limited to representing only a single relationship between the dependent and independent variables. The major advantages of LISREL are that it can estimate multiple and interrelated dependence relationships, and that it can represent unobserved concepts, or latent variables, in those relationships and account for measurement error in the estimation process.

LISREL is a technique that allows for separate relationships for each of a set of dependent variables. In its most basic sense, LISREL provides an efficient estimation technique for a series of multiple regression equations projected simultaneously. The general LISREL model consists of two sub-models: a measurement model and a structural model. The measurement model allows the researcher to use several observed variables as factors of a single unobserved independent or dependent variable. This provides the link between observed scores on measurement instruments and the underlying constructs that they are designed to measure. Using **confirmatory factor analysis (CFA)**, the researcher can evaluate the contribution of each manifest variable as well as incorporate how well the overall instrument measures the concept into the estimation of the relationships between dependent and independent variables.

The structural model is the 'path' model that defines relationships among the unobserved variables. It specifies which latent variables cause changes in the values of other latent variables in the model. The development of the structural model requires the incorporation of theory, previous experience or other guidelines to help the researcher discern which independent variables predict each dependent variable.

LISREL is usually viewed as a confirmatory rather than an exploratory procedure; however, its flexibility provides the researcher with a versatile modelling program that can be applied to a variety of research objectives. Three distinct strategies appropriate for LISREL include:

1 the strictly confirmatory strategy
2 the competing models approach
3 the model development strategy.

In the strictly confirmatory approach, the researcher specifies a single model, which is tested using LISREL's goodness-of-fit tests to assess its statistical significance. This allows the researcher to discern whether the pattern of variances and covariances in the data is consistent with the specified structural model. However, research has shown that this approach is subject to confirmation bias and tends to confirm that the model fits the data. Accordingly, other unexamined models may fit the data as well or better, and an accepted model is confirmed as being only one of several possible acceptable models.

Using the competing models strategy, the researcher may test several causal models against each other to determine which has the best fit. By examining models with different hypothetical structural relationships, the researcher comes much closer to comparing competing 'theories'. This is a much stronger test than a comparison of slight modifications of a single theory. An example is a comparison of equivalent models – alternatives that differ in proposed relationships but that have the same number of parameters and the same level of model fit.

The model development strategy differs from the strictly confirmatory and competing models approaches in that the researcher seeks to improve the model by modifying the structural and measurement models. In practice, much of the research combines confirmatory and exploratory purposes, especially in cases where theory provides only the framework for a model, which must then be empirically tested. The tested model is often found to be deficient in some way, and an alternative model is then tested based on changes suggested by the LISREL modification indexes. This is perhaps the most common approach found in the literature.

However, this approach may be problematic if the post hoc models that are confirmed do not fit new data. One way to possibly avoid this problem is for researchers to use a cross-validation approach that uses a calibration data sample to develop the model and an independent validation sample for confirmation.

Method

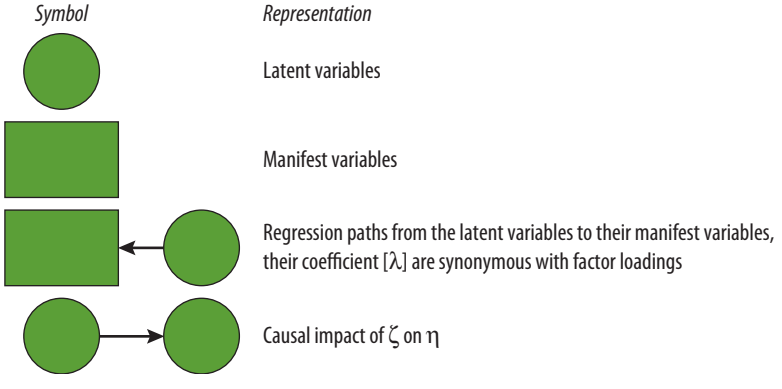
LISREL is a several-step process that begins with the development of a theoretical model based on causal relationships. In order to assume causality, the researcher must satisfy four general requirements:

1 There must be sufficient association between the two variables being considered.
2 The assumed cause must occur prior to the observed effect.
3 There must be a lack of viable alternative causal variables.
4 There must be a theoretical basis for the relationship.

Researchers must be careful to consider all key predictive variables in order to avoid **specification error**, a bias that overestimates the importance of the variables included in the model.

The second step is to construct a **path diagram** that allows the researcher to present the predictive and associative relationships among the constructs and indicators in the model. The path diagram includes various shapes, lines and notation, and researchers using LISREL must understand how the geometric symbolism depicted in the schematic models relates to the regression or matrix equations. In path diagrams, three types of arrows are used to depict all the relationships in the model. Straight arrows denote causal relationships from one variable to another; a curved arrow or line indicates a covariance between constructs; and a straight arrow with two heads shows a reciprocal relationship between constructs. Exhibit 20.10 illustrates some other key symbols frequently used in path diagrams.

Exhibit 20.10 Symbols frequently used in path diagrams.



The matrices used in LISREL equations are represented using Greek notation. A matrix is a collection of numbers written in rows and columns, and the numbers within the matrix are its elements. These elements represent the parameters in the model. A matrix with only one column but multiple rows is called a vector. Matrices are most commonly represented by uppercase Greek letters, and the elements of a matrix are denoted using lowercase Greek letters. The observed measures are indicated using Roman letters – independent variables are labelled *X* variables, and dependent variables are labelled *Y* variables. At the most, eight matrices and four vectors define a general LISREL model, although the actual number of matrices required will depend on the particular model specified.

For the general LISREL model, the measurement model is composed of two regression matrices, two variance–covariance matrices among errors of measurement, and one vector representing the endogenous factor. The structural model comprises two variance–covariance matrices (one among the exogenous factors and one among the residual errors associated with the endogenous factors) and three vectors representing the exogenous variables, endogenous variables and errors associated with the endogenous variables, respectively. A summary of these matrices and vectors is presented in Exhibit 20.11.

Exhibit 20.11 A summary of matrix and Greek notations.[9]

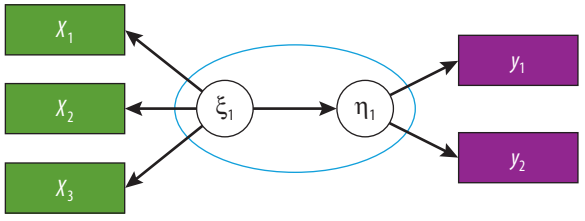
Greek letter	Full matrix	Matrix elements	Type
Measurement Model			
Lambda-*X*	ΛX	λX	Regression
Lambda-*Y*	ΛY	λY	Regression
Theta delta	Θ_δ	θ_δ	Var/cov
Theta epsilon	Θ_ε	θ_ε	Var/cov
Nu	–	V	Vector
Structural Model			
Gamma	Γ	γ	Regression
Beta	B	β	Regression
Phi	Φ	ϕ	Var/cov
Psi	Ψ	ψ	Var/cov
Xi	–	ξ	Vector
Eta	–	η	Vector
Zeta	–	ζ	Vector

In the path diagram, each of the constructs is specified as being exogenous (independent) or endogenous (dependent). Exogenous variables are not predicted by other variables, whereas endogenous variables are

predicted by other constructs. The researcher must make the distinction between exogenous and endogenous constructs with respect to each of the following:

1 the number of factors (ξ or η)
2 the number of observed variables (X or Y)
3 relationships between the observed variables and latent factors (λ_X or λ_Y)
4 factor variances and covariances (φ)
5 error variances (and possibly covariances) associated with the observed variables (Θ_ε or Θ_ε).

Exhibit 20.12 Path analysis of x and y constructs.[10]



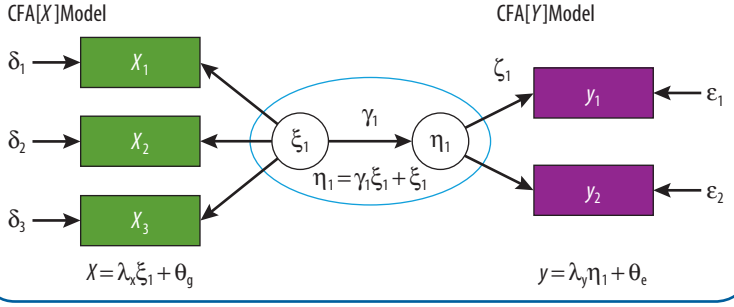
Path diagrams also require two basic assumptions. First, all causal relationships must be indicated. This requires a theoretical justification for including or excluding variables from the model. The second assumption is that the causal relationships are linear in nature. However, modifications of the LISREL equation, as in multiple regression, usually allow for the model to remain robust against this assumption. An example of a simple path analysis is shown in Exhibit 20.12.

After the path diagram is hypothesized, the third stage in LISREL is to convert the path diagram into a more formal set of structural and measurement models. This is accomplished through a set of equations that define (i) the structural equations linking the constructs, (ii) the measurement model specifying exogenous and endogenous variables, and (iii) a set of matrices indicating any hypothetical covariances among the constructs or variables. The goal for this stage is to develop a connection between the operational definitions of the constructs and the theory for the proper test.

The measurement model, commonly referred to as a confirmatory factor analysis (CFA) model, is achieved similar to exploratory factor analysis (EFA) and details on its methodology are provided in the relevant section of the chapter. The main difference between EFA and CFA is that in EFA there are no constraints on variable loading; accordingly, each variable has a loading on each factor. In CFA, the researcher specifies which variables, or indicators, define each construct. To develop the structural model, each endogenous construct becomes a dependent variable in a separate equation. Exhibit 20.13 illustrates that, relative to the structural model, there are two CFA models with a theoretical causal relationship, γ_1, between ξ_1 and η_1. δ and ε represent the measurement error associated with the observed variables, and ξ_1 is the residual error in the prediction of η_1 from ξ_1.

Next, the researcher must select the input matrix type and estimate the proposed model. At this point the researcher must test whether the data seriously violate any of LISREL's four basic assumptions: independent observations; random sampling of respondents; linearity of all relationships; and multivariate normality. Because LISREL differs from other multivariate techniques in that it only uses the variance–covariance matrix as its input data, the researcher must run diagnostic tests for violations of these assumptions using a separate statistics package, such as PRELIS, SPSS, SAS or other software.

Exhibit 20.13 The LISREL, CFA model relative to the LISREL full model.[11]



During the analysis, the researcher must select the estimation procedure used to yield the overall LISREL model. True to its name, maximum likelihood estimation (MLE) generates estimates that have the greatest probability of reproducing the observed data. MLE is by far the most common method used. This approach is advantageous over ordinary least squares (OLS) because MLE does not assume uncorrelated error terms and, accordingly, may be used for both recursive and

non-recursive models. Other estimation procedures, such as bootstrapping, simulation and jack-knifing – all of which generate samples for comparing models – are also appropriate in special circumstances.

The estimation procedure selected, as well as model misspecification error, model size and departures from normality, will all affect the sample size required for a robust model. Model misspecification error is the omission of important constructs or indicators. The researcher should increase the sample size if this is a concern. An appropriate model size is 5–10 respondents per parameter; an absolute minimum sample size is one that is greater than the number of covariances in the input data matrix. However, if the data violate the assumption of normality, we recommend a ratio of 15 respondents per parameter.

After testing the assumptions, the researcher must identify the structural model. This includes considering the size of the covariance matrix in proportion to the number of estimated coefficients. The difference between the number of covariances and the actual number of coefficients in the proposed model is the degrees of freedom, calculated as:

$$d.f. = \frac{1}{2}[(p+q)(p+q+1)] - t$$

where

p = the number of endogenous indicators
q = the number of exogenous indicators
t = the number of estimated coefficients in the proposed model.

The order condition states that the model must be just-identified or over-identified, meaning that the *d.f.* of the model must be equal to or greater than zero. The goal of the researcher is to achieve an acceptable fit with the largest number *d.f.* obtainable, which causes the model to be in its most generalizable state. In addition to meeting the order condition, the model also must meet the rank condition; that is, the researcher must algebraically determine if each parameter is uniquely estimated. A set of heuristics is available so that the researcher will not have to complete this task in its entirety.

Next, the researcher must evaluate the goodness-of-fit criteria. Goodness-of-fit tests are used to determine whether the model should or should not be rejected. If the model is not rejected, the researcher will continue the analysis and interpret the path coefficients in the model. LISREL currently provides at least 15 different goodness-of-fit measures, each of which can be categorized as one of three types of measure:

1 an absolute fit measure
2 an incremental fit measure, or
3 a parsimonious fit measure.

The type of fit index to report will be specific to the researcher's situation.

After the model is fit, the measurement model is reassessed with each construct evaluated for unidimensionality and composite reliability. Unidimensionality is an assumption for calculating reliability. Reliability measures, such as Cronbach's alpha, do not ensure unidimensionality, but they do detect whether the indicators of a construct have an acceptable fit on a single factor model. A test for construct reliability – to verify that indicators are consistent in their measurement – is not available on LISREL but can be calculated easily with the equation:

$$\text{Construct reliability} = \sum (\text{standardized loading})^2 / \sum (\text{standardized loading})^2 + \sum_{\varepsilon j}$$

In addition to the estimated coefficients, the researcher also considers the standard errors and t values for each coefficient. Because of the sensitivity of MLE with smaller sample sizes, the critical values should be conservative (a significance of either .025 or .01). And similar to multiple regression analysis, the overall R_2 is a comparative measure of fit for each LISREL equation.

As we said earlier in this section, the model can be compared with competing or nested models to find the best fit among a set of models and, if necessary, respecified to produce a model with better fit. However, at this point the researcher should be careful to evaluate only the empirical relationships – those that are not essential to the model's underlying theory. Relationships between constructs and indicators essential to the model's underlying theory should not be modified. This allows the researcher to compare several competing models with the same theoretical foundation. The researcher also can look for possible improvements by examining the residuals of the predicted

Exhibit 20.14 ***Path diagram for employee performance (exogenous variable model).***

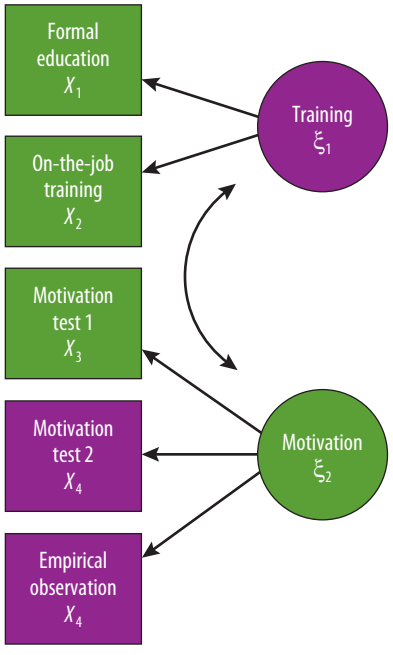
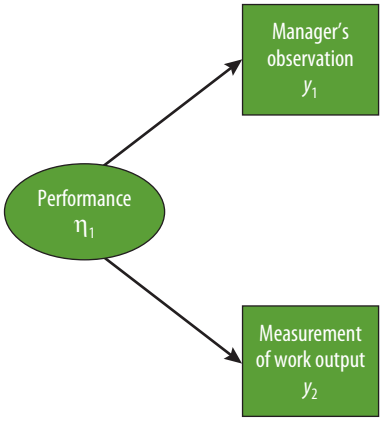


Exhibit 20.15 ***Path diagram for employee performance (endogenous variable model).***



covariance matrix. Residual values ±2.58 are considered statistically significant at the .05 level. Additionally, other tools, such as the modification index and the unexpected change parameter, can be used to assess goodness of overall model fit. However, each time modifications are made, the researcher must re-evaluate the modified model.

Example

Assume we wish to develop a model for employee performance. In conceptualizing our measurement model, we first consider a hypothetical two-factor model for employee performance with the two factors being training and motivation. These will be designated as exogenous constructs, since training and motivation are believed to have a causal effect on performance. For this example, training consists of two indicators: formal education and on-the-job training. Motivation is measured by three indicators: motivation test 1, motivation test 2 and the empirical observation of behaviour as measured on a scale. A diagrammatic representation of the exogenous model is shown in Exhibit 20.14. Here, the two-factor CFA model consists of training and motivation, with each factor measured by two and three indicators, respectively. The curved two-headed arrow denotes that training and motivation are correlated.

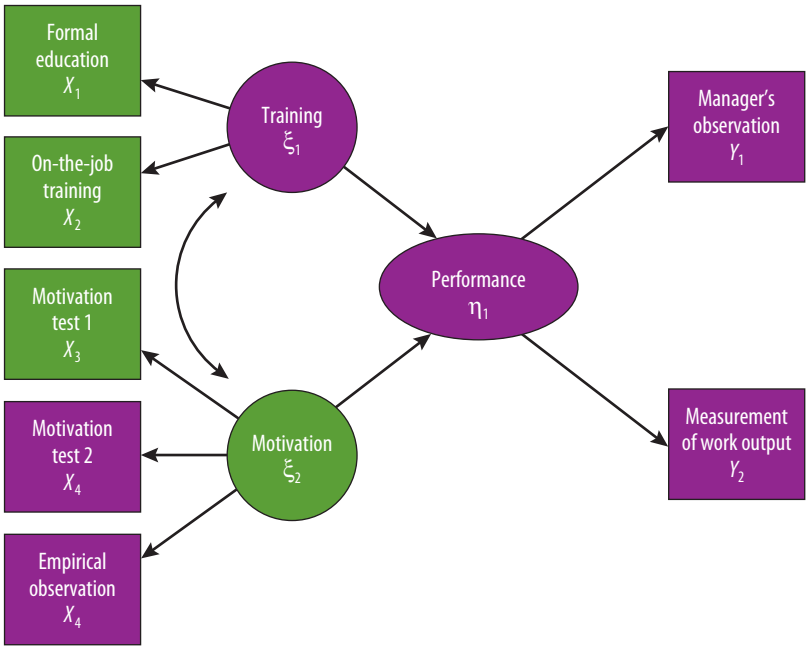
Performance is considered to be an endogenous factor, in that it is believed to be caused by motivation and training. This variable is measured by manager's observation and by the measurement of work output. The schematic representation of this CFA model is presented in Exhibit 20.15.

The full measurement model consists of a pair of CFA models identical to those developed individually. LISREL allows the researcher to test and interpret the parameters of the measurement model in exactly the same way as the parameters of the individual CFA models were tested and interpreted. The full measurement model is expressed diagrammatically in Exhibit 20.16. The structural model component would then causally relate performance to training and motivation, as well as training and motivation to their respective indicators using a system of linear structural equations. Remember that the researcher must make the distinction between the exogenous and endogenous variables, and there is no specification of causal relationships among latent variables in CFA modelling. Accordingly, the residual error (ξ) associated with the prediction of performance from motivation and training in the measurement model will be zero.

Conjoint analysis

In management research, the most common applications for **conjoint analysis** are market research and product development. Consumers buying a laptop, for example, may evaluate a set of attributes to choose the product that best meets their needs. They may consider brand, speed, price, design, educational values or capacity for work-related tasks. The attributes and their features require the buyer to make trade-offs in the final decision-making. Another application of conjoint analysis are factorial surveys (also called vignette studies). In such studies participants are confronted with hypothetical vignettes, that is, descriptions of a subject or situation along predefined dimensions. For example, in a study investigating the criteria that human resource managers use to select job applicants, you could present them with hypothetical written application letters and CVs, and then ask them

Exhibit 20.16 ***Path diagram for employee performance (full measurement model).***



what the chances are that the 'hypothetical' applicant will be invited for a job interview or selected for a job. Alternatively, you could also ask participants to compare two vignettes and ask which of the two vignettes they would prefer. Usually, participants have to assess (paired) vignettes in such a factorial survey.

Method

Conjoint analysis typically uses input from non-metric independent variables. Normally, we would use cross-classification tables to handle such data, but even multiway tables become quickly overwhelmed by the complexity. If there were three prices, three brands, three speeds, two designs, two levels of educational values and two categories for work assistance, the model would have 216 decision levels ($3 \times 3 \times 3 \times 2 \times 2 \times 2$). A choice structure this size poses enormous difficulties for respondents and analysts. Conjoint analysis solves this problem with various optimal scaling approaches, often with log-linear models, to provide researchers with reliable answers that could not be obtained otherwise.

The objective of conjoint analysis is to secure part-worths, or **utility scores**, that represent the importance of each aspect of a product or service in the subjects' overall preference ratings. Utility scores are computed from the subjects' rankings or ratings of a set of cards. Each card in the deck describes one possible configuration of combined product attributes.

The first step in a conjoint study is to select the attributes most pertinent to the decision, for example the purchase decision in the laptop example. This may require an exploratory study such as a focus group, or it could be done by an expert with a thorough knowledge of the subject investigated. The attributes selected are the independent variables, which are called **factors**. The possible values for an attribute are called factor levels. In the laptop example, the speed factor may have levels of 800 MHz, 1 GHz and 1.5 GHz. Speed, like price, approaches linear measurement characteristics since consumers typically choose higher speeds and lower prices. Other factors, like brand, are measured as discrete variables.

After selecting the factors and their levels, a computer program determines the number of product descriptions necessary to estimate the utilities. SPSS procedures ORTHOPLAN, PLANCARDS and CONJOINT build a file structure for all possible combinations, generate the subset required for testing, produce the card descriptions and

analyse results. The command structure within these procedures provides for holdout sampling, simulations and other requirements frequently used in commercial applications.[12]

Example

Watersports enthusiasts know the dangers of ultraviolet (UV) light. It fades paint and clothing, yellows surfboards, skis and sailboards, and destroys sails. More important, UV damages the skin and the eye's retina and cornea. Many consumers, however, purchase sunglasses that fail to provide adequate UV protection. Manufacturers of sunglasses for speciality markets have improved their products to such a degree that all the companies in our example advertised 100 per cent UV protection. Many other features influence trends in this market. For this example, we chose four factors from information contained in a review of sun protection products.[13]

Variable	Categories			
Brand	Sea Optics	Snow Fun Glasses	Chester	Picard
Style	A: Multiple colour choices for frame and lenses	B: Multiple colour choices for frame, limited colour choices for lenses	C: Limited colour choices for frame and lenses	
Flotation	Yes	No		
Price	€100	€72	€60	€40

This is a 4 × 3 × 2 × 4 design, or a 96-option full-concept study. The algorithm selected 16 cards to estimate the utilities for the full concept. Combinations of interest that were not selected can be estimated later from the utilities. In addition, four holdout cards were administered to subjects but evaluated separately. The cards shown in Exhibit 20.17 were administered to a small sample (n = 10). Subjects were asked to order their cards from most to least desirable. The data produced the results presented in Exhibits 20.18 and 20.19.

Exhibit 20.17 ***Concept cards for conjoint sunglasses study.***

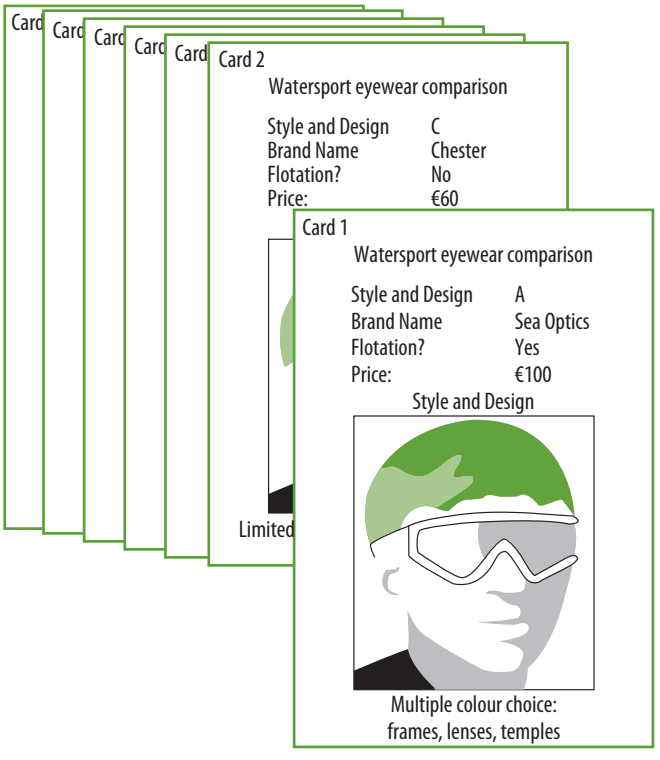


Exhibit 20.18 ***Conjoint results for subject 8, sunglasses study.***

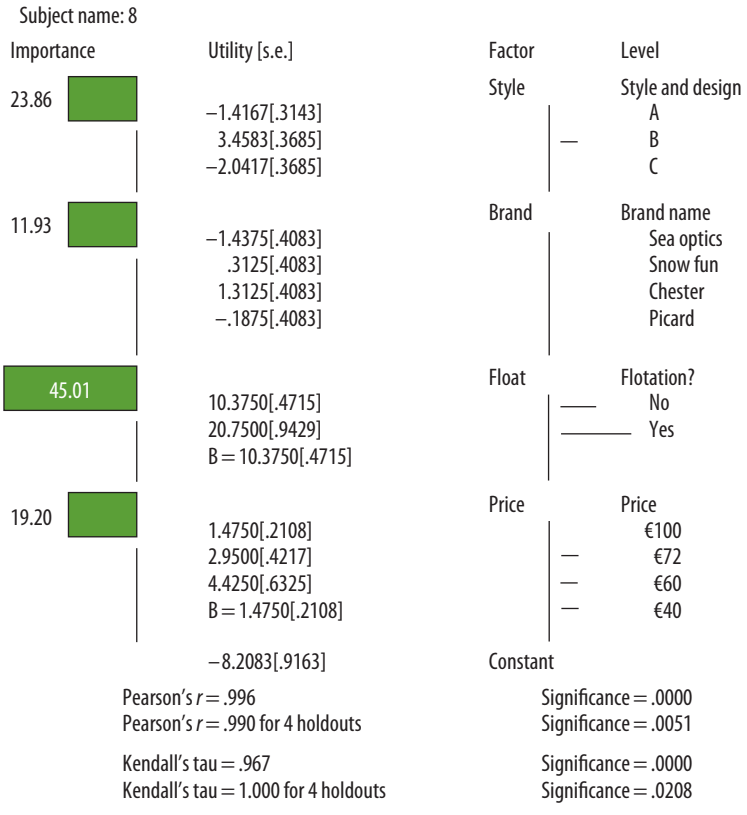


Exhibit 20.19 ***Conjoint results for sunglasses study sample.***

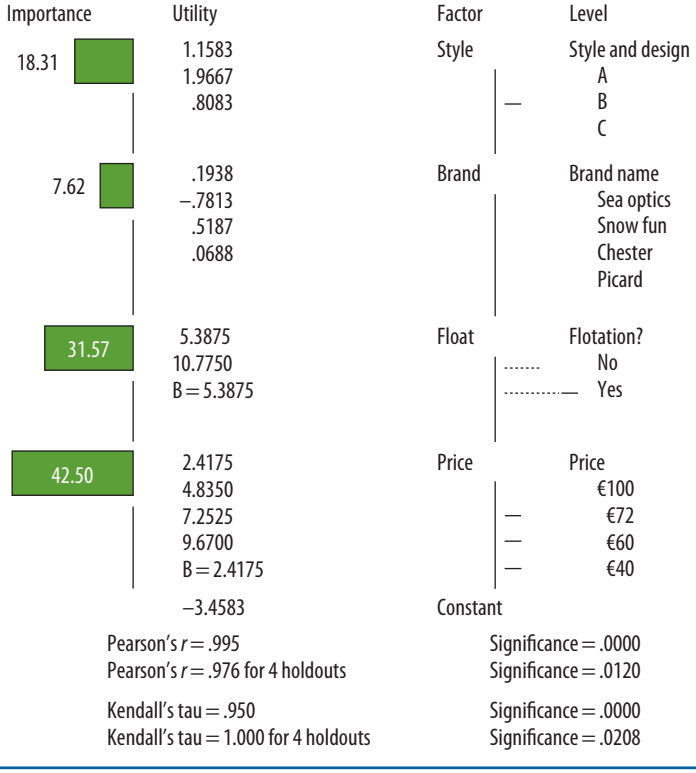


Exhibit 20.18 contains the results of the eighth subject's preferences. This individual was an avid boardsailor and flotation was the most important attribute for her, followed by style and price, and then brand. From her preferences, we can compute her total utility score:

(Style B) 3.46 + (Chester brand) 1.31 + (flotation) 20.75 + (price @ €40) 5.90 + (constant) − 8.21 = 23.21

If brand and price remain unchanged, a design that offered limited colour choices for frame and lenses (Style C) and no flotation would produce a considerably lower total utility score for this respondent. For example:

(Style C) − 2.04 + (Chester brand) 1.31 + (no float) 10.38 + (price @ €40) 5.90 + (constant) − 8.21 = 7.34

We could also calculate other combinations that would reveal the range of this individual's preferences.

Our prediction that respondents would prefer less expensive prices did not hold for the eighth respondent. She reversed herself once on price to get flotation. Other subjects also reversed once on price to trade off for other factors.

The results for the sample are presented in Exhibit 20.19. In contrast to individuals, the sample placed price first in importance, followed by flotation, style and brand. Group utilities may be calculated just as we did for the individual. At the bottom of the printout we find Pearson's r and Kendall's tau. Each was discussed in Chapter 18. In this application, they measure the relationship between observed and estimated preferences. Since holdout samples (in conjoint, regression, discriminant and other analysis methods) are not used to construct the estimating equation, the coefficients for the holdouts are often a more realistic index of the model's fit.

20.4 Interdependency techniques

Factor analysis

Factor analysis is a general term for several specific computational techniques. All have the objective of reducing many variables to a manageable number of variables that belong together and have overlapping measurement characteristics. The predictor–criterion relationship that was found in the dependence situation is replaced by a matrix of intercorrelations among several variables, none of which is viewed as being dependent on another. For example, one may have data on 100 employees with scores on six attitude scale items.

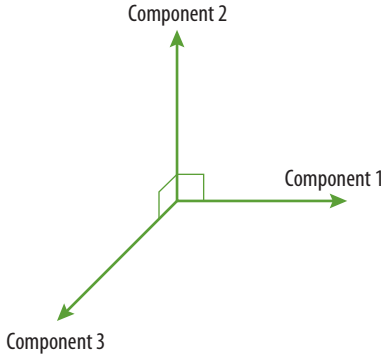
Method

Factor analysis begins with the construction of a new set of variables based on the relationships in the correlation matrix. While this can be done in a number of ways, the most frequently used approach is **principal components analysis**. This method transforms a set of variables into a new set of composite variables or principal components that are not correlated with each other. These linear combinations of variables, called factors, account for the variance in the data as a whole. The best combination makes up the first principal component and is the first factor. The second principal component is defined as the best linear combination of variables for explaining the variance not accounted for by the first factor. In turn, there may be a third, fourth and *k*th component, each being the best linear combination of variables not accounted for by the previous factors.

The process continues until all the variance is accounted for, but as a practical matter it is usually stopped after a small number of factors have been extracted. The output of a principal components analysis might look like the hypothetical data shown in Exhibit 20.20.

Numerical results from a factor study are shown in Exhibit 20.21. The values in this table are correlation coefficients between the factor and the variables (.70 is the r between variable A and factor I). These correlation coefficients are called **loadings**. Two other elements in Exhibit 20.21 need explanation. Eigenvalues are the sum of the variances of the factor values (for factor I the eigenvalue is $.70^2 + .60^2 + .50^2 + .60^2 + .60^2$). When divided by the number of variables, an eigenvalue yields an estimate of the amount of total variance explained by the factor. For example, factor I accounts for 36 per cent of the total variance. The column headed h^2 gives the **communalities**, or estimates of the variance in each variable that is explained by the two factors. With variable A, for example, the

Exhibit 20.20 ***Principal components analysis.***



Extracted components	% of variance accounted for	Cumulative variance
Component no. 1	63	63
Component no. 2	29	92
Component no. 3	8	100

Exhibit 20.21 ***Factor matrices.***

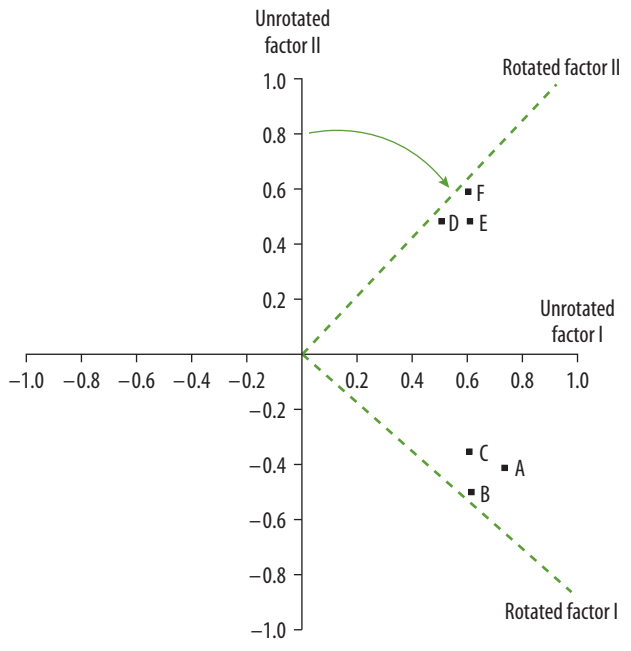
Variable	A Unrotated factors			B Rotated factors	
	I	II	h^2	I	II
A	.70	−.40	.65	.79	.15
B	.60	−.50	.61	.75	.03
C	.60	−.35	.48	.68	.10
D	.50	.50	.50	.06	.70
E	.60	.50	.61	.13	.77
F	.60	.60	.72	.07	.85
Eigenvalue	2.18	1.39			
Per cent of variance	36.30	23.20			
Cumulative per cent	36.30	59.50			

communality is $.70^2 + (-.40)^2 = .65$, indicating that 65 per cent of the variance in variable A is statistically explained in terms of factors I and II.

In this case, the unrotated factor loadings are not enlightening. What one would like to find is some pattern in which factor I would be heavily loaded (have a high r) on some variables and factor II on others. Such a condition would suggest rather 'pure' constructs underlying each factor. You attempt to secure this less ambiguous condition between factors and variables by **rotation**. This procedure can be carried out by either orthogonal or oblique methods, but only the former will be illustrated here.

To understand the rotation concept, consider that you are dealing only with simple two-dimensional rather than multidimensional space. The variables in Exhibit 20.21 can be plotted in two-dimensional space as shown in Exhibit 20.22. Two axes divide this space and the points are positioned relative to these axes. The location of these axes is arbitrary and they represent only one of an infinite number of reference frames that could be used to reproduce the matrix. As long as you do not change the intersection points and keep the axes at right angles, when an orthogonal method is used you can rotate the axes to find a better solution or position for the reference axes. 'Better' in this case means a matrix that makes the factors as pure as possible (each variable loads onto as few factors as possible). From the rotation shown in Exhibit 20.22, it can be seen that the solution is improved substantially. Using the rotated solution suggests that the measurements from six scales may be summarized by two underlying factors (see the rotated factors section in Exhibit 20.21). The interpretation of factor loadings is largely subjective. There is no way to calculate the meanings of factors; they are what one sees in them. For this reason, factor analysis

Exhibit 20.22 ***Orthogonal factor rotations.***



is largely used for exploration. One can detect patterns in latent variables, discover new concepts and reduce the amount of data by combining many items into one factor. Factor analysis is also applied to test hypotheses with confirmatory models using LISREL.

Example

Student grades make for an interesting example. The director of Hillside University's MBA programme has been reviewing grades for the first-year students and is struck by the patterns in the data. His hunch is that distinct types of people are involved in the study of management, and he decides to gather evidence for this idea.

Suppose a sample of 21 grade reports is chosen for students in the middle of the GPA range. Three steps are followed:

1 Calculate a correlation matrix between the grades for all pairs of the 10 courses for which data exist.
2 Factor-analyse the matrix by the principal components method.
3 Select a rotation procedure to clarify the factors and aid in interpretation.

Exhibit 20.23 shows a portion of the correlation matrix. These data represent correlation coefficients between the 10 courses. For example, grades secured in V1 (Financial Accounting) correlated rather well (0.56) with grades received in course V2 (Managerial Accounting). The next best correlation with V1 grades is an inverse correlation (–.44) with grades in V7 (Production).

After the correlation matrix, the extraction of components is shown in Exhibit 20.24. While the program will produce a table with as many as 10 factors, you choose, in this case, to stop the process after three factors have been extracted. Several features in this table are worth noting. Recall that the communalities indicate the amount of variance in each variable that is being 'explained' by the factors. Thus, these three factors account for about 73 per cent of the variance in grades in the financial accounting course. It should be apparent from these communality figures that some of the courses are not explained well by the factors selected.

Exhibit 20.23 ***Correlation coefficients, MBA Study Hillside University.***

Variable	Course	V1	V2	V3	V10
V1	Financial Accounting	1.00	.56	.17	–.01
V2	Managerial Accounting	.56	1.00	–.22	.06
V3	Finance	.17	–.22	1.00	.42
V4	Marketing	–.14	.05	–.48	–.10
V5	Strategic Management	–.20	–.26	–.05	–.23
V6	Organizational Behaviour	–.21	–.00	–.56	–.05
V7	Production	–.44	–.11	–.04	–.08
V8	Business Research Methods	.30	.06	.07	–.10
V9	Statistical Inference	–.05	.06	–.32	.06
V10	Quantitative Analysis	–.01	.06	.42	1.00

Exhibit 20.24 ***Factor matrix using principal factor with iterations, MBA Study Hillside University.***

Variable	Course	Factor 1	Factor 2	Factor 3	Communality
V1	Financial Accounting	.41	.71	.23	.73
V2	Managerial Accounting	.01	.53	–.16	.31
V3	Finance	.89	–.17	.37	.95
V4	Marketing	–.60	.21	.30	.49
V5	Strategic Management	.02	–.24	–.22	.11
V6	Organizational Behaviour	–.43	–.09	–.36	.32
V7	Production	–.11	–.58	–.03	.35
V8	Business Research Methods	.25	.25	–.31	.22
V9	Statistical Inference	–.43	.43	.50	.62
V10	Quantitative Analysis	.25	.04	.35	.20
Eigenvalue		1.83	1.52	.95	
Per cent of variance		18.30	15.20	9.50	
Cumulative per cent		18.30	33.50	43.00	

The Eigenvalue row in Exhibit 20.24 is a measure of the explanatory power of each factor. For example, the Eigenvalue for factor 1 is 1.83 and is computed as follows:

$$1.83 = (.41)^2 + (.01)^2 + \ldots + (.25)^2$$

The per cent of variance accounted for by each factor in Exhibit 20.24 is computed by dividing Eigenvalues by the number of variables. When this is done, one sees that the three factors account for about 43 per cent of the total variance in course grades.

In an effort to further clarify the factors, a varimax (orthogonal) rotation is used to secure the matrix shown in Exhibit 20.25. The heavy factor loadings for the three factors are also shown in this exhibit.

Exhibit 20.25 ***Varimax rotated factor matrix, MBA Study Hillside University.***

Factor 1		Factor 2		Factor 3	
Financial Accounting	.84	Finance	.90	Marketing	.65
Managerial Accounting	.53	Organizational Behaviour	–.56	Statistical Inference	.79
Production	–.54				

Variable	Course	Factor 1	Factor 2	Factor 3
V1	Financial Accounting	.84	.16	–.06
V2	Managerial Accounting	.53	–.10	.14
V3	Finance	–.01	.90	–.37
V4	Marketing	–.11	–.24	.65
V5	Strategic Management	–.13	–.14	–.27
V6	Organizational Behaviour	–.08	–.56	–.02
V7	Production	–.54	–.11	–.22
V8	Business Research Methods	.41	–.02	–.24
V9	Statistical Inference	.07	.02	.79
V10	Quantitative Analysis	–.02	.42	.09

Interpretation

The varimax rotation appears to clarify the relationship among course grades, but as pointed out earlier, the interpretation of the results is largely subjective. We might interpret the above results as showing three kinds of student, classified as the accounting, finance and marketing types. Other interpretations could be made as well.

A number of problems affect the interpretation of these results. Among the major ones are the following:

1 The sample is small and any attempt at replication might produce a different pattern of factor loadings.
2 From the same data, another number of factors rather than three can result in different patterns.
3 Even if the findings are replicated, the differences may be due to the varying influence of professors or the way they teach the courses rather than to the subject content.
4 The labels may not truly reflect the latent construct that underlies any factors we extract.

This suggests that factor analysis can be a demanding tool to use. It is powerful, but the results must be interpreted with great care.

SPSS reference

Pallant (2013) covers how to conduct a factor analysis in SPSS in Chapter 15.

Research Methods in Real Life

Big data correlations making profit

If we want to forecast buying behaviour, we usually think in terms of influencing factors that can predict whether one is likely to buy, for example, cereals or not. But the world has turned into a world of 'big data', as information technology enables us to collect and store vast amounts of data as a by-product while we shop online or use our mobile phone to call a friend. Retailers plough through these data to find patterns.

When Amazon.com started, it would base its recommendations on your previous shopping behaviour and how you rated other books; Greg Linden changed that approach. He suggested not to look at individual buyers anymore, but to look at which books sell well with other books. Nowadays, Amazon's recommendations are solely based on the correlations between products, and are believed to trigger a third of Amazon's sales. For a long time, marketers tried to understand why we buy certain products, but big data offers a new, more pragmatic approach. For years, the American superstore Walmart has recorded shopping histories along with a lot of other contextual information at the time of purchase, such as weather information. What sells well when a hurricane is approaching? Candles, batteries and bottled water are obvious and correct guesses. But would you have expected that sales of strawberry Pop-Tarts increase sevenfold? Walmart discovered this particular correlation when it looked for correlations between local weather conditions and the sales of specific items.

There are many more examples of successful data-mining, ranging from its use to create better medical therapies to improved maintenance schedules for mobile phone operators. Big data allow predictions without a dependent variable.

References and further reading

http://glinden.blogspot.de/2006/05/early-amazon-end.html. **The collected blog posts of Greg Linden, about his early days at Amazon**

http://www.nytimes.com/2004/11/14/business/yourmoney/14wal.html. ***New York Times* article about Walmart and tracking consumer buying patterns**

http://www.nytimes.com/2013/10/20/technology/to-catch-up-walmart-moves-to-amazon-turf.html. ***New York Times* article on the competition and e-commerce strategies of Walmart and Amazon**

Cluster analysis

Unlike techniques for analysing the relationships between variables, **cluster analysis** is a set of techniques for grouping similar objects or people. Originally developed as a classification device for taxonomy, its use has spread because of classification work in medicine, biology and other sciences. Its visibility in those fields and the availability of high-speed computers to carry out the extensive calculations have sped its adoption in engineering, economics, business and management studies and a host of other areas.

Cluster analysis shares some similarities with factor analysis, especially when factor analysis is applied to people (Q-analysis) instead of to variables. It differs from discriminant analysis in that discriminant analysis begins with a well-defined group composed of two or more distinct sets of characteristics in search of a set of variables to separate them. Cluster analysis starts with an undifferentiated group of people, events or objects, and attempts to reorganize them into homogeneous subgroups.

Method

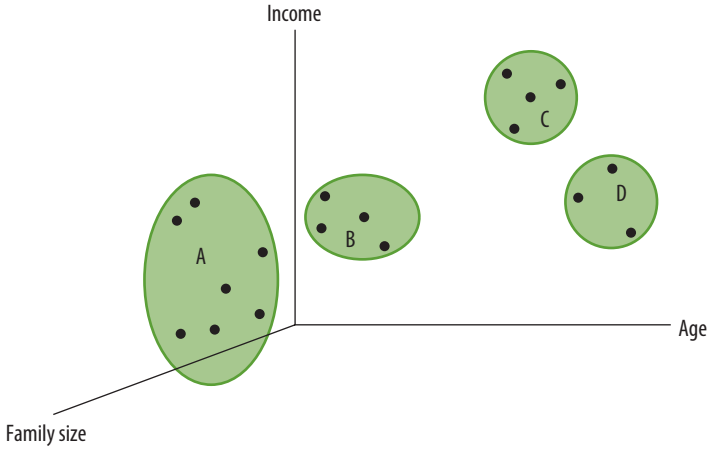
Five steps are basic to the application of most cluster studies:

1. selection of the sample to be clustered (e.g. buyers, medical patients, inventory, products, employees)
2. definition of the variables on which to measure the objects, events or people (e.g. financial status, political affiliation, market segment characteristics, symptom classes, product competition definitions, productivity attributes)
3. computation of similarities among the entities through correlation, Euclidean distances and other techniques
4. selection of mutually exclusive clusters (maximization of within-cluster similarity and between-cluster differences) or hierarchically arranged clusters
5. cluster comparison and validation.

Different clustering methods can and do produce different solutions. It is important to have enough information about the data to know when the derived groups are real and not merely imposed on the data by the method.

The example shown in Exhibit 20.26 shows a cluster analysis of individuals based on three dimensions: age, income and family size. Cluster analysis could be used to segment the car-buying population into distinct markets. For example, cluster A might be targeted as potential minivan or sport-utility vehicle buyers. The market segment represented by cluster B might be a sports and performance car segment. Clusters C and D could both be targeted as buyers of limousines, but the C cluster might be the luxury buyer. This form of clustering or a hierarchical arrangement of the clusters may be used to plan marketing campaigns and develop strategies.

***Exhibit 20.26** Cluster analysis on three dimensions.*



Example

Serious film fans find that Paris offers one of the world's best selections of films. We selected ratings from 12 cinema reviewers using sources ranging from *Le Monde* to international publications sold in Paris. The reviews reputedly influence box-office receipts, and the entertainment business takes them seriously.

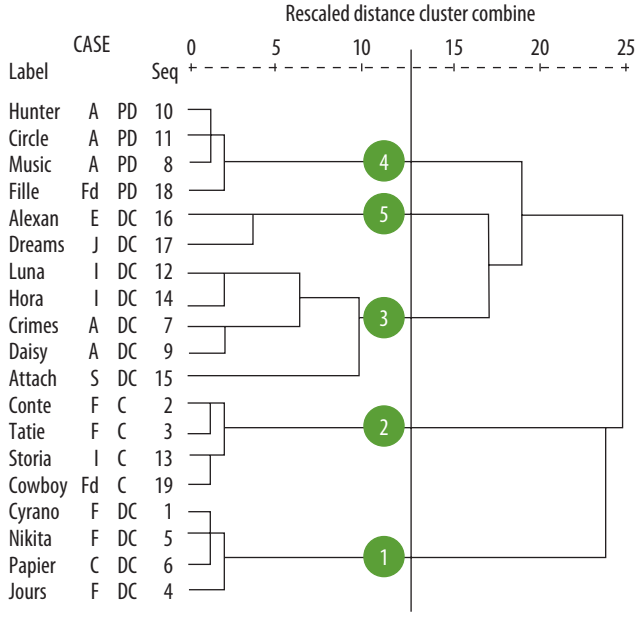
The object of this cluster example was to classify 20 films into homogeneous sub-groups. The production companies were American, Canadian, French, Italian, Spanish, Finnish, Egyptian and Japanese. Three genres of film

Exhibit 20.27 ***Film, country, genre and cluster membership.***

Film	Country	Genre	Case	Number of clusters			
				5	4	3	2
Cyrano de Bergerac	France	DramaCom	1	1	1	1	1
Il y a des Jours	France	DramaCom	4	1	1	1	1
Nikita	France	DramaCom	5	1	1	1	1
Les Noces de Papier	Canada	DramaCom	6	1	1	1	1
Leningrad Cowboys . . .	Finland	Comedy	20	2	2	2	2
Storia di Ragazzi . . .	Italy	Comedy	13	2	2	2	2
Conte de Printemps	France	Comedy	2	2	2	2	2
Tatie Danielle	France	Comedy	3	2	2	2	2
Crimes and Misdem . . .	USA	DramaCom	7	3	3	3	2
Driving Miss Daisy	USA	DramaCom	9	3	3	3	2
La Voce della Luna	Italy	DramaCom	12	3	3	3	2
Che Hora E	Italy	DramaCom	14	3	3	3	2
Attache-Moi	Spain	DramaCom	15	3	3	3	2
White Hunter Black . . .	USA	PsyDrama	10	4	4	3	2
Music Box	USA	PsyDrama	8	4	4	3	2
Dead Poets Society	USA	PsyDrama	11	4	4	3	2
La Fille aux All . . .	Finland	PsyDrama	18	4	4	3	2
Alexandrie, Encore . . .	Egypt	DramaCom	16	5	3	3	2
Dreams	Japan	DramaCom	17	5	3	3	2

were represented: comedy, dramatic comedy and psychological drama. Exhibit 20.27 shows the data by film name, country of origin and genre. The table also lists the clusters for each film using the average linkage method. This approach considers distances between all possible pairs rather than just the nearest or furthest neighbour.

Exhibit 20.28 ***Dendrogram of film study using average linkage method.***



The sequential development of the clusters and their relative distances are displayed in a diagram called a dendrogram. Exhibit 20.28 shows that the clustering procedure begins with 20 films and continues until all the films are again an undifferentiated group.

The solid vertical line shows the point at which the clustering solution best represents the data. This determination was guided by coefficients provided by the SPSS program for each stage of the procedure. Five clusters explain this dataset.

The first cluster shown in Exhibit 20.28 has three French-language films and one Canadian film, all of which are dramatic comedies. Cluster two consists of comedy films. Two French and two other European films joined at the first stage, and then these two groups came together at the second stage. Cluster three, composed of dramatic comedies, is otherwise diverse. It is made up of two American films with two Italian films adding to the group at the fourth stage. Late in the clustering process,

cluster three is completed when a Spanish film is appended. In cluster four, we find three US psychological dramas combined with a Finnish film at the second stage. In cluster five, two very different dramatic comedies are joined in the third stage.

Cluster analysis classified these productions based on reviewers' ratings. The similarities and distances are influenced by film genre and culture (as defined by the translated language).

Multidimensional scaling

Multidimensional scaling (MDS) creates a special description of a respondent's perception about a product, service or other object of interest. This often helps the business researcher to understand difficult-to-measure constructs such as product quality or desirability. In contrast to variables that can be measured directly, many constructs are perceived and cognitively mapped in different ways by individuals. With MDS, items that are perceived to be similar will fall close together in multidimensional space, and items that are perceived to be dissimilar will be further apart.

Method

We may think of three types of attribute space, each representing a multidimensional map. First, there is objective space, in which an object can be positioned in terms of its measurable attributes: its flavour, weight and nutritional value. Second, there is subjective space, where perceptions of the object's flavour, weight and nutritional value may be positioned. Objective and subjective attribute assessments may coincide, but often they do not. A comparison of the two allows us to judge how accurately an object is being perceived. Individuals may hold different perceptions of an object simultaneously, and these may be averaged to present a summary measure of perceptions. In addition, a person's perceptions may vary over time and in different circumstances; such measurements are valuable to gauge the impact of various perceptions affecting actions, such as advertising programmes.

With a third map we can describe respondents' preferences using the object's attributes. This represents their ideal; all objects close to this ideal point are interpreted as preferred by respondents to those that are more distant. Ideal points from many people can be positioned in this preference space to reveal the pattern and size of preference clusters. These can be compared to the subjective space to assess how well the preferences correspond to perception clusters. In this way, cluster analysis and MDS can be combined to map market segments and then examine products designed for those segments.

Example

We illustrate multidimensional scaling with a study of 16 companies from *BusinessWeek's* 'Executive Compensation Scoreboard'.[14] The companies chosen are from the natural resources (fuel) segment of the scoreboard. *BusinessWeek* data included executive total compensation (salary, bonus and long-term compensation for two years), shareholders' return (the year-end value based on $100 invested in corporate stock for two prior years), and the company's return on common equity (ROE) for a three-year period. We created a metric algorithm measuring the similarities among the 16 companies based on total executive compensation and the ROE. The matrix of similarities is shown in Exhibit 20.29. Higher numbers reflect the items that are more dissimilar.

If we were using respondents and producing a matrix of similarities among the perception of objects, we might obtain ordinal data. Then the matrix would contain ranks with 1 representing the most similar pair and n indicating the most dissimilar pair.

A computer program is used to analyse the data matrix and generate a spatial map.[15] The objective is to find a multidimensional spatial pattern that best reproduces the original order of the data. For example, the most similar pair (companies 3, 6) must be located in this multidimensional space closer together than any other pair. The least similar pair (companies 14, 15) must be the furthest apart. The computer program presents these relationships as a geometric configuration so all distances between pairs of points closely correspond to the original matrix.

Determining how many dimensions to use is complex. The more dimensions of space we use, the more likely the results will closely match the input data. Any set of n points can be satisfied by a configuration of $n - 1$ dimensions.

Exhibit 20.29 ***Similarities matrix of 16 companies, executive compensation.***

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
1	0															
2	3.9	0														
3	4.7	6.7	0													
4	4.4	2.8	4.7	0												
5	14.0	12.4	18.5	15.2	0											
6	4.9	6.9	0.2	4.9	18.7	0										
7	0.8	3.7	4.1	3.7	14.5	4.3	0									
8	6.0	2.1	8.5	4.0	11.8	8.7	5.8	0								
9	4.3	6.9	1.1	5.3	18.3	1.2	3.8	8.9	0							
10	8.2	4.9	8.5	4.1	15.3	8.6	7.6	3.9	9.3	0						
11	8.6	8.7	4.7	5.9	21.1	4.5	7.8	9.7	5.7	7.7	0					
12	2.2	3.7	6.9	5.5	11.8	7.1	2.8	5.5	6.5	8.5	10.5	0				
13	8.4	9.8	3.7	7.2	22.0	3.5	7.8	11.2	4.5	10.0	2.9	10.6	0			
14	12.8	13.4	8.2	10.6	25.8	8.1	12.1	14.4	9.1	12.0	4.7	14.9	4.6	0		
15	20.1	18.2	23.8	21.0	6.2	24.0	20.7	17.8	23.4	21.5	26.9	16.9	27.4	31.5	0	
16	2.6	5.2	2.1	4.0	16.5	2.3	2.0	7.2	1.9	8.0	6.3	4.8	5.8	10.3	21.7	0

Source: Similarities matrix based on data from 'Executive Compensation Scoreboard', *International BusinessWeek*, 7 May 1990, pp. 74–75.

Our aim, however, is to secure a structure that provides a good fit for the data and has the fewest dimensions. MDS is best understood using two or at most three dimensions.

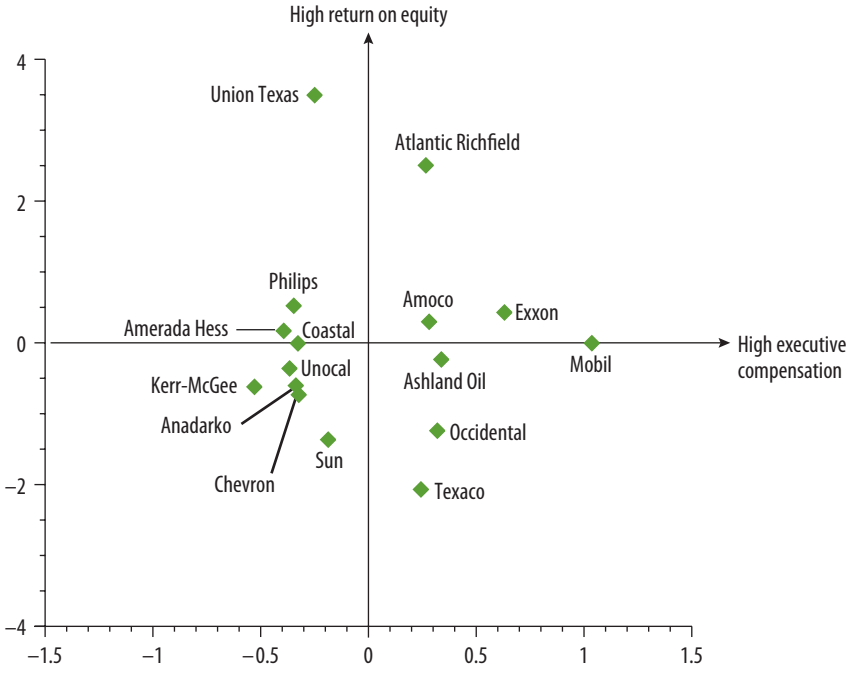
Most algorithms include the calculation of a **stress index** (S-stress or Kruskal's stress) that ranges from the worst fit (1) to the perfect fit (0). This study, for example, had a stress of .001. Another index, R^2, is interpreted as the proportion of variance of transformed data accounted for by distances in the model. A result close to 1.0 is desirable.

In the executive compensation example, we conclude that two dimensions represent an acceptable geometric configuration, as shown in Exhibit 20.30. The distance between Anadarko and Chevron (3, 6) is the shortest, while that between Texaco and Union Texas Petro Holdings (14, 15) is the longest. As with factor analysis, there is no statistical solution to the definition of the dimensions represented by the *X* and *Y* axes. The labelling is judgemental and depends on the insight of the researcher, analysis of information collected from respondents, or another basis. Respondents are sometimes asked to state the criteria that they used for judging the similarities, or they are asked to judge a specific set of criteria. In this example, the horizontal dimension approximates the total executive compensation while the vertical dimension represents return on equity.

Consistent with raw data, Union Texas and Atlantic Richfield have high ROE but compensate their executives close to the sample mean. In contrast, Exxon and Mobil generated an ROE close to the sample's average while providing higher compensation for their executives. We could hypothesize that the latter two companies may be more difficult to run – are larger and more complex – but that would need to be confirmed with another study. The clustering of companies in attribute space shows that they are perceived to be similar along the dimensions measured.

MDS is most often used to assess perceived similarities and differences among objects. Using MDS allows the researcher to understand constructs that are not directly measurable. The process provides a spatial map that shows similarities in terms of relative distances. It is best understood when limited to two or three dimensions that can be displayed graphically.

Exhibit 20.30 ***Multidimensional scaling plot of natural resource companies' return on equity and executive compensation.***



Source: Input data from *International BusinessWeek*, 7 May 1990, pp. 74–5.

Summary

1 Multivariate techniques are classified into two categories: dependency and interdependency. When a problem reveals the presence of criterion and predictor variables, we have an assumption of dependence. If the variables are interrelated without designating some dependent and others independent, then interdependence of the variables is assumed. The choice of techniques is guided by the number of dependent and independent variables involved and whether they are measured on metric or non-metric scales.

2 Multiple regression is an extension of bivariate linear regression. When a researcher is interested in explaining or predicting a metric dependent variable from a set of metric independent variables (although dummy variables may also be used), multiple regression is often selected. Regression results provide information on the statistical significance of the independent variables, the strength of association between one or more of the predictors and the criterion, and a predictive equation for future use.

3 Discriminant analysis is used to classify people or objects into groups based on several predictor variables. The groups are defined by a categorical variable with two or more values, whereas the predictors are metric. The effectiveness of the discriminant equation is based not only on its statistical significance but also on its success in correctly classifying cases to groups.

4 Multivariate analysis of variance, or MANOVA, is one of the more adaptive techniques for multivariate data. MANOVA assesses the relationship between two or more metric dependent variables and classificatory variables or factors. MANOVA is most commonly used to test differences among samples of people or objects. In contrast to ANOVA, MANOVA handles multiple dependent variables, thereby simultaneously testing all the variables and their interrelationships.

5 The LISREL technique is extremely useful in explaining causality among constructs that cannot be directly measured. LISREL has two parts, a measurement model and a structural equation model. The measurement model is used to relate the observed, recorded or measured variables to the latent variables (constructs). The structural equation model specifies causal relationships, causal effects and unexplained variance among the constructs.

6 Conjoint analysis is a technique that typically handles non-metric independent variables. Conjoint analysis allows the researcher to determine the importance of product or service attributes and the levels or features that are most desirable. Respondents provide preference data by ranking or rating cards that describe products. These data become utility weights of product characteristics by means of optimal scaling and log-linear algorithms.

7 Principal components analysis extracts uncorrelated factors that account for the largest portion of variance from an initial set of variables. Factor analysis also attempts to reduce the number of variables and discover the underlying constructs that explain the variance. A correlation matrix is used to derive a factor matrix from which the best linear combination of variables may be extracted. In many applications, the factor matrix will be rotated to simplify the factor structure.

8 Unlike techniques for analysing the relationships between variables, cluster analysis is a set of techniques for grouping similar objects or people. The cluster procedure starts with an undifferentiated group of people, events or objects, and attempts to reorganize them into homogeneous sub-groups.

9 Multidimensional scaling (MDS) is often used in conjunction with cluster analysis or conjoint analysis. It allows a respondent's perception about a product, service or other object of attitude to be described in a spatial manner. MDS helps the business researcher to understand difficult-to-measure constructs such as product quality or desirability, which are perceived and cognitively mapped in different ways by different individuals. Items judged to be similar will fall close together in multidimensional space and are revealed numerically and geometrically by spatial maps.

Discussion questions

Terms in review

1 Distinguish between multidimensional scaling, cluster analysis and factor analysis.

2 Describe the differences between dependency techniques and interdependency techniques. When would you choose a dependency technique?

Making research decisions

3 How could discriminant analysis be used to provide insight into MANOVA results where the MANOVA has one independent variable (a factor with two levels)?

4 Describe how you would create a conjoint analysis study of off-road vehicles. Restrict your brands to three, and suggest possible factors and levels. The full-concept description should not exceed 256 decision options.

5 What type of multivariate method do you recommend in each of the following cases and why?
- **a** You want to develop an estimating equation that will be used to predict which applicants will come to your university as students.
- **b** You would like to predict family income using such variables as education and stage in family life cycle.
- **c** You wish to estimate standard labour costs for manufacturing a new dress design.
- **d** You have been studying a group of successful salespeople. You have given them a number of psychological tests. You want to get some meaning from these test results.

6 Sales of a product are influenced by the salesperson's level of education and gender, as well as consumer income, ethnicity and wealth.
- **a** Formulate this statement as a multiple regression model (form only, without parameter estimation).
- **b** Specify dummy variables.
- **c** If the effects of consumer income and wealth are not additive alone, and an interaction is expected, specify a new variable to test for the interaction.

7 What multivariate technique would you use to analyse each of the following problems? Explain your choice.
- **a** Employee job satisfaction (high, normal, low) and employee success (0–2 promotions, 3–5 promotions, 5+ promotions) are to be studied in three different departments of a company.
- **b** Consumers making a brand choice decision between three brands of coffee are influenced by their own income levels and the extent of advertising of the brands.
- **c** Consumer choice of colour in fabrics is largely dependent on ethnicity, income levels and the temperature of the geographical area. There is detailed area-wide demographic data available on income levels, ethnicity and population, as well as the weather bureau's historical data on temperature. How would you identify geographical areas for selling dark-coloured fabric? You have sample data for 200 randomly selected consumers: their fabric colour choice, income, ethnicity and the average temperature of the area where they live.

From concept to practice

8 An analyst sought to predict the annual sales for a home-furnishing manufacturer using the following predictor variables:

X_1 = Marriages during the year.
X_2 = Housing starts during the year.
X_3 = Annual disposable personal income.
X_4 = Time trend (first year = 1, second year = 2, and so on).

Using data for 24 years, the analyst calculated the following estimating equation:

$$Y = 49.85 - .068X_1 + .036X_2 + 1.22X_3 + 20.54X_4$$

The analyst also calculated an $R^2 = .92$ and a standard error of estimate of 11.9. Interpret the above equation and statistics.

9 A researcher was given the assignment of predicting which of three actions would be taken by the 280 employees in a Surrey plant that was going to be sold to its employees. The alternatives were to:
- **a** take severance pay and leave the company
- **b** stay with the new company and give up severance pay
- **c** take a transfer to the plant in Leeds.

The researcher gathered data on employee opinions, inspected personnel files and the like, and then did a discriminant analysis. Later, when the results were in, she found the results listed below. How successful was the researcher's analysis?

Actual decision	Predicted decision		
	A	B	C
A	80	5	12
B	14	60	14
C	10	15	70

10 You are working with a consulting group that has a new project for the Landsend School System. The school system of this district has individuals with purchasing, service and maintenance responsibilities. They were asked to evaluate the vendor/distribution channels of products that the county purchases.

The evaluations were on a 10-point metric scale for the following variables.

- Delivery speed – amount of time for delivery once the order has been confirmed.
- Price level – level of price charged by the product suppliers.
- Price flexibility – perceived willingness to negotiate on price.
- Manufacturer's image – manufacturer or supplier's image.
- Overall service – level of service necessary to preserve a satisfactory relationship between buyer and supplier.
- Sales force – overall image of the manufacturer's sales representatives.
- Product quality – perceived quality of a particular product.

(The data can be viewed at www.mcgraw-hill.co.uk/textbooks/blumberg.)

Your task is to complete an exploratory factor analysis on the survey data. The purpose for the consulting group is twofold: (i) to identify the underlying dimensions of these data, and (ii) to create a new set of variables for inclusion into subsequent assessments of the vendor/distribution channels.

Issues to consider in your analysis are as follows.

a Methodology: (i) desirability of principal components versus principal axis factoring; (ii) decisions on criteria for number of factors to extract; (iii) rotation of the factors; (iv) factor loading significance; and (v) interpretation of the rotated matrix.

b Prepare a report summarizing your findings and interpreting your results.

11 The data file 'venture_capital_europe' (available on the website) consists of country level data for the period 1995 to 2003. For each country and year we have data how well the countries venture capital sector did that year and what kind of strategy companies followed.

a Use regression models to find out what are the effects of certain strategies on market performance.

b What are the problems of using simple multiple regression on this dataset.

c Panel regression models would be more appropriate, if you are familiar with those models, recalculate the coefficients you have produced in a. What are the differences?

d In the dataset a couple of data points are missing, what could you do?

Recommended further reading

Cohen, Jacob, Patricia Cohen, Stephen West and Leona Aiken, *Applied Multiple Regression / Correlation Analysis for the Behavioral Sciences* (3rd edn). Mahwah, NJ: Lawrence Erlbaum Associates, 2002. A widely used reference guide for all issues related to various forms of regression analysis.

***Sage Series in Quantitative Applications in the Social Sciences*. Thousand Oaks, CA: Sage Publishing.** This monograph series includes papers on most multivariate methods.

Schumaker, Randall A. and Richard G. Lomax, *A Beginner's Guide to Structural Equation Modelling* (3rd edn). New York: Routledge, 2010. A popular introduction to structural equation models.

Get started with understanding statistical techniques!

When you have read this chapter, log on to the Online Learning Centre website at *www.mcgraw-hill.co.uk/textbooks/blumberg* to explore chapter-by-chapter test questions, additional case studies, a glossary and more online study tools for *Business Research Methods*.

Notes

1 Jagdish N. Sheth (ed.), *Multivariate Methods for Market and Survey Research*. Chicago, IL: American Marketing Association, 1977, p. 3.

2 William Schneider, 'Opinion outlook', *National Journal* (July 1985).

3 Benson Shapiro, 'Price reliance: existence and sources', *Journal of Marketing Research*, August 1973, pp. 286–9.

4 For a discussion of path analysis, see Elazar J. Pedhazur, *Regression in Behavioral Research: Explanation and Prediction* (2nd edn). New York: Holt, Rinehart & Winston, 1982, Chapter 15.

5 Fred Kerlinger, *Foundations of Behavioral Research* (3rd edn). New York: Holt, Rinehart & Winston, 1986, p. 562.

6 Joseph F. Hair Jr., Rolph E. Anderson, Ronald L. Tatham and William C. Black, *Multivariate Data Analysis with Readings*. New York: Macmillan, 1992, pp. 153–81.

7 This section is based on the SPSS procedure MANOVA, described in Marija J. Norusis/SPSS, Inc., SPSS *Advanced Statistics Users Guide*. Chicago: SPSS, Inc., 1990, pp. 71–104.

8 This section was prepared by Jeff Stevens, School of Public Administration, Florida Atlantic University. It is based on J. Hair, R. Anderson, R. Tatham and W. Black, *Multivariate Data Analysis with Readings* (5th edn). Upper Saddle River, NJ: Prentice Hall, 1998; J. ScottLong, *Covariance Structure Models: An Introduction to LISREL*. Thousand Oaks, CA: Sage Publishing, 1984; J. ScottLong, *Confirmatory Factor Analysis: A Preface to LISREL*. Thousand Oaks, CA: Sage Publishing, 1983; and Barbara M. Byrne, *A Primer of LISREL: Basic Applications and Programming for Confirmatory Factor Analytic Models*. New York: Springer-Verlag, 1989.

9 From B.M. Byrne, *A Primer of LISREL: Basic Applications and Programming for Confirmatory Factor Analytic Models*. New York: Springer-Verlag, 1989, p. 8.

10 Ibid., p. 9.

11 Ibid., p. 9.

12 SPSS, Inc., *SPSS Categories*. Chicago: SPSS, Inc., 1990.

13 Product specifications adapted from Lewis Rothlein, 'A guide to sun protection essentials', *Wind Rider* (June 1990), pp. 95–103.

14 'Executive Compensation Scoreboard', *International BusinessWeek*, 7 May 1990, pp. 74–5.

15 See the ALSCAL procedure in Marija J. Norusis/SPSS, Inc., *SPSS Base System User's Guide*. Chicago: SPSS, Inc., 1990, pp. 397–416.